Analysis of Large Genomic Data *in Silico* -- the EPIC-Norfolk Study of Obesity

Jing Hua Zhao
MRC Epidemiology Unit
Institute of Metabolic Science
Box 285
Addenbrooke's Hospital, Hills Road
Cambridge CB2 0QQ
United Kingdom

Tel: +44 (0)1223 769165
email: jinghua.zhao@mrc-epid.cam.ac.uk
Web: http://www.mrc-epid.cam.ac.uk/~jinghua.zhao/

29/4/2008

Introduction

This is a suite of programs to conduct genomewide association analyses for both binary and continuous (or possibly other type) traits. Due to the enormous amount of data from such studies and growing complexity involved it is necessary to perform the analysis using established software. Nevertheless, it is quite possible that even with system such as SAS, the usual computer specification is insufficient. The suite approaches this by splitting the whole data into a number of manageable pieces approximately of equal sizes. The master program then combines results from individual runs. Information about SAS is available from http://www.sas.com. Occasionally we also refer to Stata (http://www.stata.com) and R (http://www.r-project.org). As the later stage of the analysis was heavily involved with imputing and analysing SNPs based on the HapMap (http://www.hapmap.org), relevant programs have been written. Codes for Illumina 317 GeneChips have been added to those for Affymetrix 500K GeneChips for the same set of individuals.

Methods

The bulk of the analysis is quite standard, including quality control including Hardy-Weinberg equilibrium tests, allele coding assuming dominant, recessive or additive models, and regression analyses of continuous and binary traits. The programs also give descriptive statistics associated with all SNPs, e.g. counts and mean, standard errors associated with each genotype, and gene annotation.

The prototype for this suite was an analysis of 250,000-SNP Perlegen set for 400 normal controls from a cancer study, which turned to be manageable under x86 Linux systems. However, it became impossible with Affymetrix 500k GeneChips and several thousands of individuals and therefore this suite is developed and tested.

Input

As it is, this suite accepts phenotypic data, genotypic data, physical map and gene annotation information from Affymetrix. While these files use a tabular format, the genotype data is in the form of .gz compressed file to be piped into SAS. The annotation was added when assembling results.

Map/phenotypic information is available from the current directory, while genotypes could be from another location.

The suite was originally developed following the data format adopted by the Wellcome Trust Case-Control Consortium (WTCCC) as described from http://www.wtccc.org.uk/info/data_formats.shtml but with alterations later on.

Running the programs

The suite is modular, such that components are written to accomplish individual tasks for easy maintenance. There is a master program called *go.sas* to do things systematically. As one may be limited by the amount of disk space allocated, it is advised to run the driver program as follows,

sas -work /scratch go &

provided that large space is available from /scratch. Surprisingly, -*noterminal* option is also required when running *Affy500k.sas*

It is more useful in that data on a particular chromosome are split into different batches to minimise memory requirement, which has the benefit that runs restart from where they stop.

Outputs

The output for chromosome *i* is as follows:

SAS Program	Description
code.sas	allele information (cma&i)
hwe.sas	Hardy-Weinberg equilibrium tests (caf&i,cms&i)
desc.sas	Summary statistics (cmeans&i,cfreq&i)
bt.sas	chisquares (cpm&i) and OR with CI (clpm&i)
qt.sas	anova(canova&i), parameter estimates (cparms&i)

These are located at */data/genetics/scratch* by default, but can be changed to other location by replacing the definition in *setup.sas* or overwriting it after *setup.sas* is included. In case the system runs out of memory, it would be helpful to set the number of splits (ns) to a larger value in *go.sas*. The analysis of Chromosome X data is done separately, for it is of interest to produce descriptive statistics by gender. There are two types of coding, i.e., additive and dominant, and in the additive case males are coded as 0 and 2 whereas females are coded as 0, 1 and 2. However, for completeness Hardy-Weinberg equilibrium (HWE) tests are also performed in women (sex=2) only for chromosome X.

Note the model data set used was from a case-cohort study of obesity, and body mass index (BMI) was also done in cohort samples (cohort="1").

As the analysis is quite computer-intensive, it is necessary to use *nohup* command from Linux, which allows output to non-tty. Otherwise SAS often stops with signalling hang-up.

There are only three places in *setup.sas* to be changed for Windows, i.e., *obesity*, *scratch*, *home*=. It is also more convenient to run SAS under batch mode, (e.g. *sas.bat*) and *gzip* is required in the search path. However, it is possible to run these programs from display management system (DMS), namely *dm 'x cd c:\epic5k*'.

Annotation is accomplished via *Affy500k.sas* which processes files from Affymetrix and generate additional information.

Before a summary file containing most of the information is generated, the call rates need to be calculated with *calls.sas*. Note minor change is required for file specifications. Other information includes allele frequency information from the British 1958 Birth Cohort.

A program called *sum.sas* has been written to tally results from individual components. This can be used as follows,

```
sas -work /scratch sum.sas &
```

One can modify *sum.sas* -- it is very comprehensive as it is. The validity check has been done with specific SNPs within *go.sas* by adding where statement.

As obese cases are defined according to BMI, it also makes sense to examine BMI as a quantitative trait in a truncated regression model. This can be achieved with PROC QLIM from SAS/ETS (or PROC LIFEREG). An example macro (*qlim*) is also provided. It leaves the reader for ordinary least squares (OLS) regression analysis. Nevertheless, Box-Cox regression or regressor transformation is also possible with PROC QLIM. Closely related is meta-analysis from several samples whose prototype can be seen (*meta.sas*). However, the implementation perhaps has its advantage over Stata (*meta.do*) and R (*meta.R*) only when more sophisticated models are sought.

Additional analyses

Utility programs have been written for the following software from elsewhere:

HaploView	LD and association testing (http://www.broad.mit.edu/mpg/haploview)
EIGENSTRAT	Principal components analysis for stratification (http://genepath.med.harvard.edu/~reich/Software.htm)
GTOOL	Transforming sets of genotype data for **IMPUTE** and **SNPTEST**
IMPUTE	Genotype imputation
SNPTEST	Analysis of output from IMPUTE (http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html)

Reference

Zhao JH, Luan JA, Tan Q, Loos R, Wareham N. Analysis of Large Genomic Data *in Silico* -- the EPIC-Norfolk Study of Obesity. DS Huang, L Heutte, M Long (Eds). Communications in

Computer and Information Sciences. ICIC 2007 2, 781-790, 2007

Supplementary notes

S1. An extra plate

A peculiarity associated with the sample code is the availability of an extra plate, whose files are under directory 25125. This is handled gracefully using pipe command (see *data.sas*). Because the files are named such that 01 is 1, SAS macro function *%eval* is invoked to drop the prefixed 0.

S2. Rotation of case-cohort genotypes

A further complication was described as follows,

"*Also, it should be noted that for any preliminary analysis, Sequenom identified a 180 degree plate rotation, 15575, during genotyping of the Affy500Ks. These were split plates, so OBs (obesity cases) and OBCs* (controls) *changed cohorts during the rotation. I've attached a mapping file which reassigns the wells appropriately (15575A1 is actually 15575H12, etc.), we're currently correcting this at our end. I plan to send out a QC update next week.*"

The mapping file was given as rotate180 and the issue has been resolved using SNP rs9939609 from chromosome 16 but with absolute generality. The associate SAS program is *rotate.sas* which is compatible with other codes for the whole genomewide analysis.

S3. Issues in relationship to IMPUTE, GTOOL and SNPTEST

The posted from Sanger Institute in our analysis have been reversed for the positive strand, which has the effect that to comply with the files from the imputation program we used (**IMPUTE**), we have to reorient according to Build 35 strand information, which is achieved by *TRANSLATE* function. It requires meticulous care for genotype imputation by IMPUTE, for there are discrepancy between three versions of map information (version 1 of IMPUTE, the Sanger and an updated version). In particular, the following SNPs require special attention:

SNP name	Strand	Alleles	Sanger alleles	IMPUTE version 1
rs2664070	+	AG	CT	-
rs2429893	+	CT	AG	-
rs4886982	-	CT	CT	+
rs10967532	-	AG	AG	+
rs3960362	-	CT	CT	+

Therefore it turns out the strand information as it is in the updated version is wrong for IMPUTE. The *impute.sas* deals with them separately and to keep in line with IMPUTE by noting the fact that this program only accepts allele labels in alphabetic order. Nevertheless there are other differences which will bring in further uncertainty. It thus appears to use *impute_fwd.sas* is more sensible. However, the Probe_Set_Ids should be kept as the Sanger map does not contain strand information.

Simple descriptive statistics from comparison of the observed versus imputed data are helpful. These are shown with codes such as *wtccc.sas* used for cross-checking with the WTCCC summary tables. Program *agree.sas* is written with inputs courtesy of Tim Frayling, Hana Lango in the Type-2 Diabetes group within WTCCC.

While the imputed data were analysed with **SNPTEST**, there were problems to subset individuals and therefore a C program *getline.c* and an AWK program *idex.awk* were written. It turned out that the latter would be considerably faster. One can obtain the imputed genotypes for a subset of individuals in advance to accommodate the various analyses, but an alternative would be to use **GTOOL**. Even so, AWK is useful since it is rather handy to insert the missing proportions from **SNPTEST** for an input file. Furthermore, when the software suite was developed, there were problems with covariate adjustment so separate files were generated from SAS to be used by a bash script *snptest.sh*. Now it can be simplified but very clear for the collaborative and consortium work we have been involved otherwise. The SED program could well be replaced by AWK with its *gsub* function but they are kept as they were. Although their tasks can be performed with more powerful languages such as Perl but they are much lighter.

S4. Analysis using wide format

Data in standard, tabular format may have appeal since it occupies less disk space and may allow for faster data access as well. It also allows for more flexible merging of phenotypic information. There can be two ways to achieve this, i.e., individual by SNPs, or SNPs by individuals (similar to HapMap and also IMPUTE), the latter is likely to be of value when large number of SNPs are involved. Codes are available for both. There might be difficulty to convert the merged data using these formats so codes have been written to allow for SNP-wise analysis. Please check *impute.sas, impute_fdw.sas, xpose.sas* and *vars.sas* for details.

S5. Code optimisation

Although codes are designed to fit the problem, in particular the long (skinny) format data, it is generally slow to perform analysis with new phenotype(s). The data generated from S4 provide the basis for code optimisation. To facilitate this it is a good idea to avoid PROC TRANSPOSE, so a separate program *vite.sas* has been written. This could turn out to be the key for a great success (labelled in French)! It is possible to convert the long and the wide formats data in SAS freely and possibly at faster speed but the house-keeping would be more involved without PROC TRANSPOSE. Since this is required only once even with the long format data, so we can do it once for all, that is, to have data in wide format, and then use a dedicate program *encode.sas*. As *vite.sas* and *encode.sas* have somewhat unusual use of SAS system functions and macro facilities, they are more demanding for less sophisticated users. Furthermore, *encode.sas* does not require pre-existing numeric codes and one can run regression analysis based on the data it generates. It is able to include non-SNP data such as individual IDs, phenotypes, and with the option to be in wide or long format and use files generated from one format in the other. The coded values in the wide format are character-type, so it is useful to have data in long format, which is inline with the earlier approaches.

Appendix A complete list of programs

Main analysis

setup.sas	to provide definitions of working directories, macros, etc.
trait.sas	phenotype processing
map.sas	map information
data.sas	program for raw genotypes
exclude.sas	program for generating exclusion list
code.sas	allele coding (additive, dominant, recessive)
pg.sas	phenotype-genotype merging
split.sas	data partitioning program
rotate.sas	rotate of samples in a plate
fto.sas	data extraction for FTO gene
ds2text.sas	a macro to create ASCII file (see also *eigen.sas*)
desc.sas	summary statistics
hwe.sas	Hardy-Weinberg equilibrium test
xhwe.sas	Hardy-Weinberg equilibrium test for chromosome X
allele.sas	data processing for chromosome X
go.sas	the master program for BMI and obesity case-control analysis
call.sas	call rates for SNPs
qlim.sas	truncated variable regression
bt.sas	analysis of binary trait
qt.sas	direct regression analysis without data partitioning
meta.sas	meta-analysis
metaex.sas	an example of meta-analysis with PROC MIXED
meta.do	meta-analysis (Stata program)
meta.R	meta-analysis (R program)
fdr.sas	false discovery rate (an example program)
sum.sas	summary of the main results
merge.sas	pure data generation

qqplot.sas	Chi-squared Q-Q plot
Affy500k.sas	annotation from Affymetrix 500k GeneChips
Affy500snps.sas	to read map from Sanger with strand information
bc58.sas	data from BC 1958
partition.sas	iterative merging of variables
tidy.sas	to empty data
select.sas	selection of sample and SNP ids for the refine imputation
checkv1.sas	check of strand information in IMPUTE/version1.tgz and Build 35
strand.sas	comparison of strands between map from Sanger and an update version
vars.sas	to work on individual SNPs rather than long, skinny data
xpose.sas	generation of wide format data to allow for flexible merging
vite.sas	optimised analysis based on data in wide format
encode.sas	faster allele coding based on data in wide format
sas.bat	MS-DOS batch file to start SAS
gzip.exe	MS-DOS program for generating/decompressing *.gz* file and for SAS pipe command as used in *data.sas*

awk.exe	AWK program for MS-DOS (http://www.cs.bell-labs.com/who/bwk/awk95.exe)
sed.exe	SED program for MS-DOS (http://sed.sourceforge.net/#download)

HaploView

ld.sas	data preparation

EIGENSTRAT

eigen.sas	data preparation

GTOOL/IMPUTE/SNPTEST and WTCCC data

gtool.sh	bash shell driver script for GTOOL
gtool.subs	bash shell script to be called by *gtool.sh*
impute.sas	data preparation for genotype IMPUTation
impute_fwd.sas	data preparation assuming genotypes all forward strand
impute_v1.sas	to combine information from IMPUTE *version1.tgz*
getline.c	selection of individuals according to list (optional)
strand.seds	script to put all negative strand to positive for IMPUTE
impute.sh	bash script to run IMPUTE
impute.subs	make script to be called by *impute.sh/make*
epic5k.sh	make script running EPIC-Norfolk data
epic5k_fwd.sh	make script running EPIC-Norfolk data assuming all forward strands
30.sh	bash script calling *epic5k.sh* with 30 partitions
idex.awk	awk program to be used by *snptest.subs*
idex.sh	bash script to prepare for cohort data and Affy500k SNPs
idex.do	Stata program to generate *trait.raw* for *snptest.subs*
idex.sas	SAS equivalent to *idex.do*
idex2.sas	as above for updated GIANT analyses (named convention and latest SNPTEST)

snptest.sas	analysis of IMPUTEed genotypes (see also snptest.do)
snptest.do	a Stata program similar to *snptest.sas*
snptest.sh	bash script to perform analysis using SNPTEST
snptest2.sh	as above for updated GIANT analyses
snptest.subs	Subroutine to be called by *snptest.sh*
assemble.sas	combination of outputs from SNPTEST
assemble.awk	AWK program to be called by assemble.sas
assemble.sh	bash script to call *assemble.sas* without using PROC IMPORT
assemble2,3.sh	as above for updated GIANT analyses with and without SAS
Affy500k.sh	bash script to perform analysis of Affy500k data
Affy500k.subs	Subroutine to be called by *Affy500k.sh*
agree.sas	agreement between imputed and observed regression p values
wtccc.sas	correlation of p values from observed and imputed genotypes in WTCCC

Illumina 317 GeneChips

These are available from the Illumina directory.